CLAIMS

1. A method for measuring the degree of coherence of the arrangement of nodes in a hierarchy comprising the steps of:

a) receiving a hierarchy of nodes;

b) receiving a plurality of training cases that are filed under said nodes; and

c) responsive thereto for determining a measure of coherence, for at least one node that has a local environment, by evaluating the training cases under the node with respect to the training cases in the local environment of the node.

2. The method of claim 1 wherein the step of determining a measure of coherence includes the steps of

determining, for the subtree at the node, the number of the training cases and the average prevalence of each feature in the training cases;

determining, for the local environment of the node, the number of the training cases and the average prevalence of each feature in the training cases;

determining predictive features that distinguish the subtree of the current node from the local environment of the node; and

generating a coherence value for the current node based on the average prevalence of at least one predictive feature.

3. The method of claim 2 further comprising the steps of

determining, for each said predictive feature, the degree of uniformity of the prevalence of the predictive feature among the children subtrees of the node;

and wherein the step of generating a coherence value for the current node is based on said degree of uniformity and the average prevalence of at least one predictive feature.

4. The method of claim 1 wherein the hierarchy of nodes includes a topic hierarchy; wherein the nodes are topics; and wherein the training cases includes one of labeled documents and feature vectors assigned to the topics.

5. The method of claim 2 wherein the predictive features includes one of words, multi-word phrases, noun phrases, document length, file extension type, other parameters related to documents, and a combination thereof.

6. The method of claim 2 wherein the step of determining the predictive features includes the step of

computing one of information-gain metrics, mutual-information metrics, Chi-Squared, Fisher's Exact Test, lift, odds-ratio, word frequency among documents, word frequency among all words in all documents, and a combination thereof.

7. The method of claim 3 wherein the step of selecting features that are uniformly common includes

the step of computing one of the metrics cosine-similarity, projection, and Chi-Squared between the average feature prevalence vector and the vector of training case counts across the subtopics of the current node.

8. The method of claim 3 wherein the step of generating a hierarchical coherence number includes the step of

generating a hierarchical coherence number by computing the average prevalence of the predictive feature with the greatest degree of uniformity.

9. The method of claim 3 wherein the step of generating a hierarchical coherence number includes the step of

generating a hierarchical coherence number by computing a weighted-average of the average prevalence of at least two features that are selected as both predictive and uniform.

10. The method of claim 9 wherein the step of generating a hierarchical coherence number includes the step of

generating a hierarchical coherence number by computing a weighted-average of the average prevalence of the top k most prevalent features that are selected as both predictive and uniform, wherein k is a predetermined positive integer.

11. The method of claim 9 wherein the weighted-average employs as the weighting schedule one of the negative exponential function exp(-I) and the inverse rank function (1/I), where I is the ordered rank of the most prevalent features that are selected as both predictive and uniform.

12. The method of claim 3 wherein the step of generating a hierarchical coherence number includes the step of

generating a hierarchical coherence number by computing the average value of the average prevalence of the top k most prevalent features that are selected as both predictive and uniform, wherein k is a predetermined positive integer.

13. The method of claim 2 wherein the step of generating a hierarchical coherence number includes the step of

generating a hierarchical coherence number by employing a maximum, over all predictive features, of a projection between the average feature prevalence vector and the vector of training case counts across the subtopics of the current node .

14. The method of claim 2 wherein the step of generating a hierarchical coherence number includes the step of

generating a hierarchical coherence number by employing a maximum average prevalence of the predictive features.

15. The method of claim 1 further comprising the step of:

assigning an aggregate-coherence value to a node in the hierarchy, based on an aggregation function of said determined measure of coherence over said node and of descendants of said node.

16. The method of claim 15 wherein the aggregation function includes one of a sum, average, weighted-average, minimum function, and maximum function.

17. The method of claim 1 further comprising the step of:

using the coherence values of one or more nodes to modify the structure of the hierarchy to improve the coherence of the hierarchy.

18. The method of claim 1 further comprising the step of:

using the coherence values of one or more nodes to guide the selection of training cases for an automated classifier.

19. The method of claim 1 further comprising the step of:

using the coherence values of one or more nodes to select a suitable classification technology to be employed to automatically classify items in the hierarchy.

20. An apparatus for measuring the degree of coherence of at least one considered node that has associated therewith a subtree and a local environment in a hierarchy comprising:

a) a training case counter for determining the number of training cases under the subtree and the number of training cases for the local environment;

b) an average prevalence determination unit for determining for at least one feature the average prevalence under the subtree and the average prevalence for the local environment; and

c) a predictive feature determination unit for determining the set of predictive features that distinguish training cases of the subtree from documents of the local environment; and

d) a coherence assignment unit for generating a coherence metric number for each considered node based on at least one predictive feature.

21. The apparatus of claim 20 further comprising:

a_5) a subtopic uniformity determination unit for determining the uniformity of the distribution of said predictive features among the children subtopics of the considered node;

wherein the coherence assignment unit generates a coherence metric number based on at least one predictive feature that is determined to be uniformly distributed among said children subtopics.

22. A system for measuring the degree of coherence of nodes in a topic hierarchy comprising:

a) a coherence analyzer unit for receiving the topic hierarchy and a set of labeled training cases and responsive thereto for determining, for at least one current node under consideration, a measure of coherence by evaluating the training cases and at least one

feature under the local environment of the current node and by evaluating the training cases and at least one feature under the subtree of the node under consideration.

23. The system of claim 22 further comprising:

b) a user interface presentation unit coupled to the coherence analyzer unit for displaying a measure of coherence for one or more current nodes under consideration.

24. The system of claim 22 further comprising:

b) feature extractor coupled to the coherence analyzer for receiving a set of labeled documents and at least one feature guideline and responsive thereto for generating the set of labeled feature vectors.

25. The system of claim 22 wherein the coherence analyzer further comprises:

a_1) a training case counter for determining the number of training cases under each node subtree;

a_2) an average prevalence determination unit for determining the average prevalence for at least one feature under each node subtree; and

a_3) a predictive feature determination unit for determining predictive features under each node subtree; and

a_4) a coherence assignment unit for generating coherence metric number based on at least one predictive feature.

26. The system of claim 25 wherein the coherence analyzer further comprises:

a_5) a subtopic uniformity determination unit for determining the degree of uniformity in the distribution of one or more said predictive features among the children of the current node;

wherein the coherence assignment unit generates a coherence metric number based on at least one uniform predictive feature.

27. A method for measuring the degree of coherence for one or more current nodes in a hierarchy comprising the steps of:

a) receiving a hierarchy and the training cases filed into said hierarchy;

b) determining a list of predictive features that distinguish documents of the current node's sub-tree from those in the current node's local environment;

c) assigning a coherence value to the current node based on the list of predictive features and based on one or more of their degree of predictiveness, their the degree of prevalence, and their degree of uniformity, wherein the degree of uniformity reflects how evenly distributed said predictive features are among the children subtrees of the current node based on the training cases under each child subtree.